PCA による複数判別変数の選択

小河原史絵　(指導教員：吉田裕亮)

1　はじめに

我々は日常生活の様々な場面で, いろいろな情報に基づいて判断を行っている. しかし, 個々人の持っている情報はきわめて範囲が広く, 複雑であるため, あるひとつの要因だけで物事を判断するのは不安である. 従って多変量のデータから判別に影響のある要因を知ることは重要となる. そこで本研究では, PCA を用いて必要な情報を 2 次元に縮約し, マハラノビスの距離を用いて, 2 群に分ける. さらにその誤判別率を用いることで, 判別に寄与する要因を見つけ出す手法を提案する.

2　主成分分析 (PCA)

主成分分析 (以下 PCA) とは, 多次元データの総合力や特性といった情報量を, それ自体はあまり少なくせずとも, より低い次元に縮約させる方法である.

X を $n \times k$ のデータ行列とし, X の縦成分 (変数) ごとに平均と標準偏差を求めて標準化し, その行列を X_0 とする. そして, 正定値行列である相関行列 R を求める.

$$R = \frac{1}{n} X_0 {}^t X_0$$

さらに, 固有方程式を解き, k 個の固有値と, 各々の固有値に対応する固有ベクトル T' を求める. 標準化されたデータ行列 X_0 を,

$$X_0 T' = X^*$$

と変換し, X^* の第 1 列目を第 1 主成分, X^* の第 2 列目を第 2 主成分と呼ぶ.

本研究では, この第 1, 2 主成分に縮約し, 誤判別率を求めるために用いる. また, 優固有値・固有ベクトルを求める手法として, 比較的簡単な手順で求められる累乗法を用いた.

3　累乗法

適当な単位ベクトル $\overrightarrow{u}$ を選び, 規格化し行列 A に掛ける. 再び規格化した A に掛けることを何度も繰り返すことにより, A の絶対値最大の固有値に属する固有ベクトルに収束する.

A が正定値行列であり, 固有値 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$ が全て異なるならば, 固有ベクトル $\overrightarrow{u_1}, \overrightarrow{u_2}, \ldots, \overrightarrow{u_n}$ は互いに直交し,

$$A = \lambda_1 \overrightarrow{u_1}^t \overrightarrow{u_1} + \lambda_2 \overrightarrow{u_2}^t \overrightarrow{u_2} + \ldots + \lambda_n \overrightarrow{u_n}^t \overrightarrow{u_n}$$

が成り立つ. 従って, 最大固有値 λ_1 と固有ベクトル $\overrightarrow{u_1}$ が見つかったならば, $A - \lambda_1 \overrightarrow{u_1}^t \overrightarrow{u_1}$ に累乗法を再び適用すれば, 次に絶対値の大きい A の固有値・固有ベクトルが得られる. 以降, 望むだけの優固有値・固有ベクトルも同様である.

4　マハラノビス距離

多変量の相関に基づき, 算出される距離としてマハラノビス距離がある. 2 次元データで各クラス ($j = 1, 2$) の平均が $\mu_j = {}^t(\mu_{1j}, \mu_{2j})$ で表される各 j 群のマハラノビス距離 D_j は, で与えられる.

$$D_j = \sqrt{{}^t(x - \mu_j)\Sigma_j^{-1}(x - \mu_j)}$$

ここで, $\sum_j^{-1}$ は j 群の分散共分散行列の逆行列である.

5　提案手法

まず, 複数の変数を持ち, 2 群に分けられた多変量データが与えられたとする. 変数をいくつか取り除いたデータに対し主成分分析を施し, 可視化可能な 2 次元に縮約する. 縮約された合成変数にマハラノビス距離を用いて, 2 群判別を行う. ここで, 本来のグループと違うグループであると判断されたものを数え上げ, 誤判別率とする.

取り除く変数が 1 つの場合は, 昨年の卒業研究 [3] で行われているため, 本研究では主に 2 複数変数を取り除いて行うことにする.

6　数値実験

5 変量で 2 群に分かれたデータを 1000 個用意する. うち 500 個は変数 X_0, X_1 が $N(-1, 1)$, 変数 X_2, X_3, X_4 が $N(0, 1)$ に従い, 残りの 500 個は変数 X_0, X_1 が $N(7, 1)$, 変数 X_2, X_3, X_4 が $N(0, 1)$ に従うものとする.

まず, 1 変数を取り除いたデータに主成分分析を施し, 誤判別率を求める. 結果は各々の誤判別率がなく, 取り除く変数が 1 変数では判別への影響がない.

次に, 2 変数ずつ取り除き, 同様に誤判別率を求める. 結果は以下のように表される,

取り除いた変数	構成変数	誤判別率
X_0, X_1	X_2, X_3, X_4	47.4%
X_0, X_2	X_1, X_3, X_4	7.4%
X_1, X_2	X_3, X_4, X_5	6.6%

ここで, その他は 0.0% である.

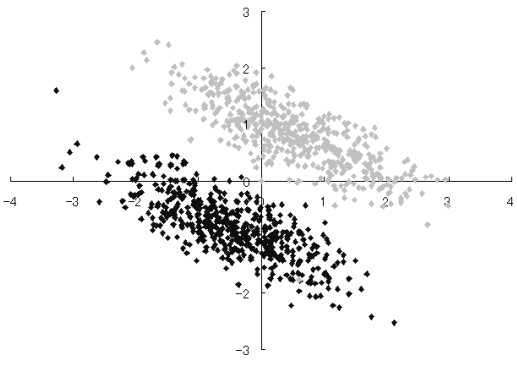

図 1: 結果

これにより, 2 変数 X_0, X_1 を取り除くと, 誤判別率が上がることが分かる. よって, X_0, X_1 の組み合わせ

が特に影響の大きい判別変数とみなされる.

7　実データへの応用

7.1　健康診断データ

実データへの応用として, 脂質代謝判定に影響のある成分の選択に当てはめてみる.

まず, 脂質代謝が正常とみなされた人の中から 435 人, 正常でないとみなされた人の中から 435 人をランダムに抽出し, 計 870 人のデータを用意する. これらは年齢, BMI, 総コレステロール, ヘモグロビン, クレアチニンの 5 つの変数 (成分) を持つ. また, これらは血液検査なしで入手可能なデータを選んだ.

7.2　推定結果

全 5 成分で主成分分析を施すと, 誤判別率は 19.2 % であった.

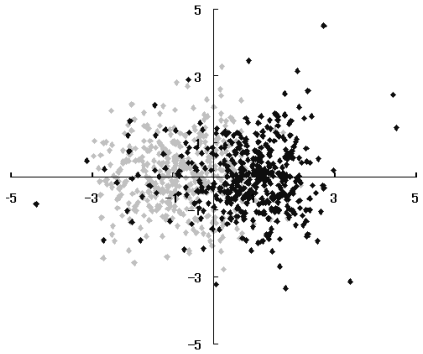

図 2: 5 成分での結果

ここから 1 変数ずつ取り除き, 同様に誤判別率を求めると, 以下のようになる.

年齢	18.0%
BMI	18.3%
総コレステロール	26.9%
ヘモグロビン	18.7%
クレアニチン	18.9%

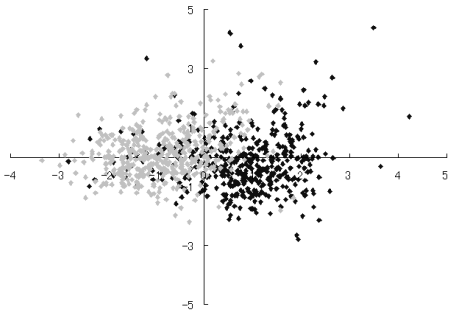

図 3: 総コレステロールを取り除いた結果

総コレステロールは影響が大きいといえるが, 他の成分はほぼ同じようになる.

さらに, 2 変数を取り除き, 誤判別率を求めると, 以下のようになる.

年齢, BMI	18.0%
年齢, 総コレステロール	38.2%
年齢, ヘモグロビン	20.7%
年齢, クレアチニン	18.2%
BMI, 総コレステロール	29.5%
BMI, ヘモグロビン	16.4%
BMI, クレアチニン	17.0%
総コレステロール, ヘモグロビン	32.5%
総コレステロール, クレアチニン	27.0%
ヘモグロビン, クレアチニン	17.7%

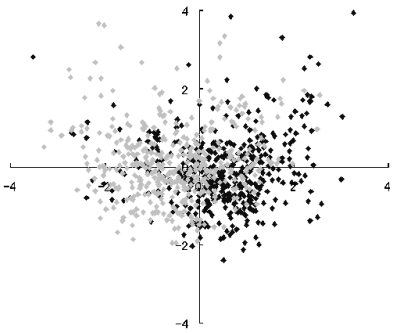

図 4: 年齢, 総コレステロールを取り除いた結果

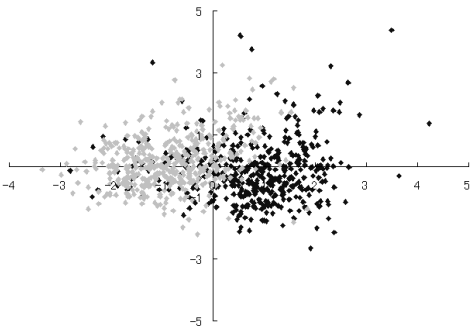

図 5: 総コレステロールとヘモグロビンを取り除いた結果

これにより, 総コレステロールと年齢, 総コレステロールとヘモグロビンの組み合わせの誤判別率が他と比べると大きいことが分かり, 判別への寄与が大きいといえる.

8　まとめ

本研究の手法を用いると, 多次元の変数の中から, 判別への寄与が大きい複数変数の組み合わせも見つけ出すことが可能といえる.

参考文献

[1] 鈴木義一郎, 情報量基準による統計解析入門, 講談社 (1995)

[2] 竹村彰通, 谷口正信, 統計学の基礎　, 岩波書店 (2003)

[3] 仁平智子“ PCA と判別分析を用いた判別変数の推定 ”, お茶の水女子大学卒業研究会要旨集, pp53−54, (2008)